SPECTRAL WARPING

AND

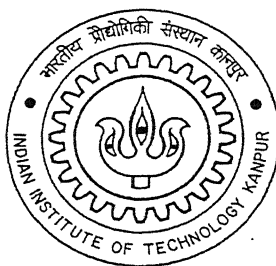# NOISE REDUCTION IN ASR SYSTEMS

A Thesis Submitted

in Partial Fulfillment of the Requirements

for the Degree of

Master of Technology

by

Rajesh Sharma

to the

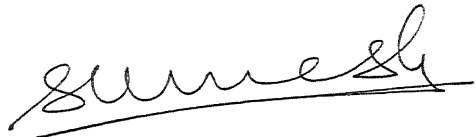## DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

April 2002

CERTIFICATE

It is certified that the work contained in the thesis entitled *"Spectral Warping and Noise Reduction in ASR Systems"* by *Rajesh Sharma* has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

(Dr. S. Umesh)

April 2002

Associate Professor,

Department of Electrical Engineering,

Indian Institute of Technology,

Kanpur-208016

Acknowledgements

At the very outset, I would like to express my heartfelt gratitude to that Supreme Being for being benevolent to me, as He has always been to all of His Manifestations, and for making me what i am.

I am heartily thankful to my thesis supervisor Dr. S. Umesh for giving me a chance to work under him and for his indispensable guidance during my thesis period. I thank him for his constant encouragement, open discussions and for being generous in bestowing me with his knowledge. His always inspiring positive attitude, informal atmosphere in the lab and his ever smiling face are unforgettable.

I would also like to thank all of my instructors at IIT, Kanpur for their invaluable guidance. I am highly thankful to my lab mate Rohit for his invaluable help during experimentation and scrupulous scrutinizing discussions. I am too much thankful to My lab mate and dear friend Bharath for enthusiastically taking part in academic or non academic discussions initiated by me at any place. The moments spent with Sir, Rohit and Bharath are really unforgettable. I am also thankful to Shafi for his help at various times

I thank all of my friends specially SachinN, Pankaj, SachinS, Shashi, Ketan Klt, Anil, Mishra, Desh Raj and my dear friend Brainy Bharath and other DSP classmates (Murali & Harsha) for making my stay at IIT, Kanpur joyous and memorable forever. These friends are very precious to me.

At last i dedicate this thesis to my beloved parents whose inspiration, sacrifices and life time efforts have helped me to come upto this point. I am also thankful to all the other family members for their generous help which i can never forget.

Rajesh Sharma

Abstract

In this thesis we have worked on two main problems faced by Automatic Speech Recognition systems namely, speaker variability and background noise. The problem of speaker variability has been investigated using non-linear spectral warping functions for speaker normalization. We have proposed a model for the warping function and the parameters of this warping function have been estimated from speech data. This warping function has been compared with the more commonly used log warping and Mel warping functions.

In the study of robustness to background noise, we have compared the recognition performance of WOSA (weighted overlapped segment averaging) and Mel Filter Bank methods of feature computation under various SNRs. We have also proposed a method for local estimation of the noise PSD that can be used in spectral subtraction. We present results comparing the recognition accuracies of the conventional method i.e. without any pre-processing of noisy speech with our proposed method of noise compensation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the first people to take a serious look at developing Automatic Speech Recognition (ASR) technology was Alexander Graham Bell. As early as the 1870s, Bell wanted to invent a machine (called a "phonoautograph") that could transcribe spoken words into written text. The purpose of the machine was to overcome the isolation and barriers to communication he saw in the deaf community. Although Bell was unsuccessful in the invention of the phonoautograph, it was while working on it that he had the inspiration for what would become the telephone.

1.1   ASR Problems

Today, ASR systems are revolutionizing the world as they are being applied to many everyday applications. But still there are some challenging problems that prevent the widespread deployment of ASR systems.

An ideal ASR system is supposed to work for variety of speakers i.e. it should be

speaker independent and it should also work in all environmental conditions i.e. it should be robust to noise and channel variations. But today's ASR systems are still susceptible to these two problems and a lot of research is directed towards developing systems that are robust to these variations.

ASR systems are divided into two main parts. The first part is known as "Front End Signal Processing" and deals with the feature extraction of speech. The most popular features are the Mel-filter bank cepstral coefficients (MFCC). The second part or the "Back End" deals with pattern classification, pattern matching and postprocessor. Nowadays the pattern classification and matching is based on Hidden Markov models (HMM). In our current study we have worked on the first part i.e. Front End Signal Processing.

There are different types of variabilities which corrupt the speech[12] e.g. spectral tilt, input level, physiological differences and additive noise. Though each one of them is a serious problem, the last two problems still pose a big challenge. Physiological differences creep due to differences in vocal tract shapes and sizes and additive noise is an environmental effect.

The physiological differences manifest as differences in spectra for the same enunciated sounds by different speakers. These differences can be mitigated by transforming these spectra into some other domain so that all the spectra for the same enunciation by different speakers look similar. This transformation is known as spectral warping. The main aim of universal warping methods is to transform sound spectra into another

domain (warped domain) so that all spectra for same enunciation by different speakers become shifted version of each other. Now this shift factor can be removed by taking magnitude of Fourier Transform of shifted spectra as it is shift invariant. This transformation is known as universal warping function. If the warping function is log then this whole idea becomes Scale Transform [2]. Though many warping methods have been proposed, the problem has not yet been satisfactorily solved. The first part of current study is centered around this theme of obtaining improved warping function.

Apart from above mentioned problem, various methods have been devised for noise robustness also. The seriousness of this problem can be judged from few lines of a recent issue of the magazine "Speech Technology" .

"One of the barriers to the implementation of speech technology is, of course, noise. No matter how good the recognition engine, without proper noise cancellation and filtering the process can be likened to trying to be heard in a thunderstorm"

For additive noise compensation, in ASR systems, researchers are mainly working in two domains, first is the model domain where they try to prepare noise based models and second is the feature domain where they try to separate noise from speech features. In later approach the usual method of noise compensation is to pass signal through a filter which suppress the noise. This is further divided into two different approaches, first is multichannel based noise cancellation i.e. adaptive noise cancellation [3], which makes use of a reference input derived from one or more sensors located at points in the noise field. This input is filtered and subtracted from a primary input containing both

signal and noise. Second method is also based on spectral subtraction [4, 5] but it makes use of only one channel. This method tries to estimate the noise from those portions of utterance where speech activity is not present assuming that noise remains stationary. For estimation of noise, in later methods, few of the methods makes use of voice activity detector (VAD) and few are based on minimum statistics[11]. All of these methods have their own advantages and disadvantages e.g. the design of noise robust VAD itself is a big problem. Similarily there are problems with other methods also. Moreover one common problem with all these methods is the difference between assumptions about the noise statistical characteristics and actual environmental conditions. Motivated by these noise related problems, we have studied the noise compensation using the idea of cross correlation.

1.2 Thesis Organization

In this thesis we worked on above two problems i.e. universal warping function for speaker normalization and the noise compensation for speech recognition. In the second chapter of this thesis we will describe the piecewise beta warping function found solely on the basis of speech data and its closeness to other warping methods like Mel and log Warping has been studied.

In the third chapter we have carried out the comparison of WOSA and Mel-filter bank methods in the context of noise robust feature extraction and discussed them in detail.

The fourth chapter is mainly devoted to the problem of noise compensation. In this chapter we have proposed a cross-correlation based noise reduction method and its performance has been compared with baseline i.e. recognition accuracy for noisy speech without pre-processing. The method makes use of the general assumptions of uncorrelatedness of noise with signal.

Finally in fifth chapter we have discussed the results and also talked about the scope of future work in this method.

Chapter 2

Spectral Warping Functions

The concept of differences in the vocal tract length in speakers leading to the major source of variability in speech is well established. This has lead to the "scale relationship" between the speakers. The motivation of speaker normalization has resulted in the application of scale invariant transforms viz. Scale Transform [7] ,Fourier–Mellin Transform [8]. in deriving the speech features [2]. The basic idea is to warp a pair of mutually scaled spectra such that in the warped domain they are shifted versions of one another. By taking the magnitude of fourier transform of these shifted versions, we get identical spectra in the warped domain.

2.1   Scale Transform

Briefly, the scale transform of a function, $X(f)$ is given by,

$$D_X(c) = \int\limits_0^\infty X(f)\, \frac{e^{-j2\pi c \ln f}}{\sqrt{f}}\, df \qquad (2.1)$$

and inversely,

$$X(f) = \int_{-\infty}^{\infty} D_X(c) \frac{e^{j2\pi c \ln f}}{\sqrt{f}} \, dc \quad \forall \ f \geq 0 \tag{2.2}$$

A basic property of the scale transform is that the magnitude of the scale transform of a function, $X(f)$ and its normalized scaled version, $\sqrt{\alpha}X(\alpha f)$, are equal (note that $0 < \alpha < 1$ corresponds to dilation, while $1 < \alpha < \infty$ corresponds to compression), since

$$\begin{aligned} D_X^\alpha(c) &= \int_0^{\infty} \sqrt{\alpha}X(\alpha f) \frac{e^{-j2\pi c \ln f}}{\sqrt{f}} \, df \\ &= e^{j2\pi c \ln \alpha} D_X(c) \end{aligned} \tag{2.3}$$

Eq.(2.3) shows that the scale transform of $\sqrt{\alpha}X(\alpha f)$ is same as that of $X(f)$ except for a linear phase, which disappears by taking magnitude on both sides of Eq. (2.3), i.e.,

$$|D_X^\alpha(c)| = |D_X(c)| \tag{2.4}$$

Thus, considering two speakers who are scaled versions of each other, α being their characteristic scale factor, Eq. (2.4) shows that in the scale transform domain, both the speakers look alike, as the speaker dependent term that appears in phase is nullified by taking magnitude. Thus, there is no need to explicitly calculate the speaker specific scaling constant. The scale transform may also be calculated as the fourier transform of the function $X(e^f)e^{\frac{f}{2}}$ i.e.,

$$D_X(c) = \int_{-\infty}^{\infty} X(e^f)e^{\frac{f}{2}} e^{-j2\pi c f} \, df \tag{2.5}$$

It is to be noted that as a result of log-warping, i.e., forming $X(e^f)$, the speaker specific scale constant, α, is purely a function of translation parameter in the log-warped domain.

2.2 Frequency Warping Function

Consider two speakers $\mathcal{A}$ and $\mathcal{B}$ related by

$$S_\mathcal{A}(f) = S_\mathcal{B}(\alpha_{\mathcal{AB}} f) \qquad (2.6)$$

where $S(.)$ denotes the spectral envelopes and $\alpha_{\mathcal{AB}}$ is the scale factor of the subject speaker $\mathcal{B}$, with respect to the reference speaker, $\mathcal{A}$. This would be the case if uniform scaling was true. Consider the warping function $f = e^v$ which is applied to all speakers. This is log-warping of the frequency axis. In the log-warped domain, the scaling factor $\alpha_{\mathcal{AB}}$ appears as a translation factor.

$$
\begin{aligned}
S_a(v) &= S_\mathcal{A}(f = e^v) = S_\mathcal{B}(\alpha_{\mathcal{AB}} e^v) \\
&= S_\mathcal{B}(e^{v + \ln \alpha_{\mathcal{AB}}}) = S_b(v + \ln \alpha_{\mathcal{AB}})
\end{aligned}
\qquad (2.7)
$$

Note the use of lower-case subscripts that denote the spectra or functions in the warped domain or v domain. Thus, in the log-warped domain, the warped spectra are shifted versions of one another. The magnitude of their fourier transform leads to scale invariance.

$$|F(S_\mathcal{A}(v))| = |F(S_\mathcal{B}(v + \ln \alpha_{\mathcal{AB}}))| \qquad (2.8)$$

where F(.) represents the fourier transform operator. Here exponential sampling denotes linear scaling of the frequency axis, which is realized as equal sampling in log-domain. But studies shows that the scale factor, α_{AB} is indeed formant dependent and context dependent, resulting in non-uniform scaling. In such a case, the relation between spectral envelopes of two speakers can be modelled as

$$S_{\mathcal{A}}(f) = S_{\mathcal{B}}(\alpha_{\mathcal{A}\mathcal{B}}(f)f) \tag{2.9}$$

where $\alpha_{AB}(f)$ is a frequency-dependent, non-uniform scaling factor. Analogous to uniform scaling, our goal is to find a transformation, $f = z(v)$ such that

$$
\begin{aligned}
S_{\mathrm{a}}(v) &= S_{\mathcal{A}}(f = z(v)) \\
&= S_{\mathcal{B}}(\alpha_{\mathcal{A}\mathcal{B}}(f)f) = S_{\mathrm{b}}(v + \varsigma_{\mathcal{A}\mathcal{B}})
\end{aligned} \tag{2.10}
$$

where $\varsigma_{\mathcal{A}\mathcal{B}}$ is dependent only on the speakers $\mathcal{A}$ and $\mathcal{B}$, and is independent of frequency. Now, our aim is to find the function $z(.)$, which warps the spectra, thus making them shifted versions in the warped domain.

The non-linearity $\alpha_{\mathcal{A}\mathcal{B}}(f)$ has been modelled in many parametric forms by different persons .We have modelled it as

$$f' = \alpha_{\mathcal{A}\mathcal{B}}(f)f = \alpha^{\beta(f)}f \tag{2.11}$$

where α is the subject's scale factor, with respect to a reference speaker, which is speaker specific and frequency independent and $\beta(f)$ is only frequency dependent and is independent of speaker. $\beta(f)$ captures the non-linearity in scale factor. Eq.(2.11)

can be modified as

$$\log(f') = \beta(f)\log(\alpha) + \log(f) \tag{2.12}$$

$$\frac{\log(f')}{\beta(f)} = \log(\alpha) + \frac{\log(f)}{\beta(f)} \tag{2.13}$$

Define

$$\nu = W(f) = \frac{\log(f)}{\beta(f)} \tag{2.14}$$

Assuming $\beta(f') \simeq \beta(f)$, we get

$$\nu' = \nu + \log(\alpha) = \nu + constant shift \tag{2.15}$$

where ν is the warped domain and $W(f)$ is the frequency warping function. Eq. (2.15) shows that the spectra in the warped domain are translated versions of one another. The magnitude of the fourier transform of these warped spectral patterns are invariant to translations, leading to scale invariant features, of real speech signals. Since finding the exact form of this warping function is difficult, we discretized the computation of the warping function [9, 10].

2.3   Numerical Computation of the Warping Function

2.3.1   Discrete-Implementation of Warping Function

We now obtain a relationship between f and v at a discrete set of frequencies. Let us divide the frequency axis into N logarithmically equi-spaced regions. In each region, let

us assume that the spectral envelopes of any two speakers are scaled versions of each other. So, for i^{th} frequency region, $f \in [L_i, U_i)$, we have

$$S_{\mathcal{A}}(f) = S_{\mathcal{B}}(\alpha_{\mathcal{AB}}^{(i)} f) \, , \; L_i \leq f < U_i \tag{2.16}$$

where $\alpha_{\mathcal{AB}}^{(i)}$ is the scale factor for i^{th} frequency region, L_i and U_i are the lower and upper frequency boundaries of ith region respectively and $1 \leq i \leq N$. Define

$$\alpha_{\mathcal{AB}}^{(i)} = \alpha_{\mathcal{AB}}^{\beta_i} \tag{2.17}$$

which assumes that the frequency dependency is present in the parameter β and $\alpha_{\mathcal{AB}}$ is only speaker dependent (independent of frequency). We need to compute $S_b(v = \log(f))$ for $v \in [\log(L_i), \log(U_i))$. Let us discretize the computation of $S_b(v)$ at M_i equally spaced intervals in the region $\log(L_i)$ to $\log(U_i)$. Let

$$\Delta v_i = \frac{\log(U_i) - \log(L_i)}{M_i} \tag{2.18}$$

Then the uniformly spaced samples in the ith frequency region in v domain are $S_b \left(m_i \Delta v_i + \log(L_i) \right)$ for $m_i = 0, 1, \cdots, (M_i - 1)$. Uniformly sampling $S_a(v)$ at Δv_i spacing in the ith frequency region results in

$$S_a \left(m_i \Delta v_i + \log(L_i) \right) = S_b \left(m_i \Delta v_i + \log(\alpha_{\mathcal{AB}}^{\beta_i}) + \log(L_i) \right) \tag{2.19}$$

Eq. (2.19) can be rewritten as

$$S_a(m_i \Delta v_i + \log(L_i)) = S_b \left(\left(m_i + \frac{\beta_i \log(\alpha_{\mathcal{AB}})}{\Delta v_i} \right) \Delta v_i + \log(L_i) \right) \tag{2.20}$$

It can be seen that the two functions differ by a translation factor $\frac{\beta_i \log(\alpha_{\mathcal{AB}})}{\Delta v_i}$ in the ith frequency region. Since we define the warped envelopes to be translated versions of

one another, over the entire range of interest, we require the following condition to be satisfied between any two frequency regions i and j.

$$\frac{\beta_i \log(\alpha_{AB})}{\Delta v_i} = \frac{\beta_j \log(\alpha_{AB})}{\Delta v_j} = \frac{1}{\Delta \lambda} \tag{2.21}$$

where λ is a new domain where the scaled spectra appear as shifted versions of one another. From Eq.(2.18), we have

$$\Delta v_i M_i = \Delta v_j M_j \tag{2.22}$$

as $\log\left(\frac{U_i}{L_i}\right) = \log\left(\frac{U_j}{L_j}\right)$, thus resulting in $\beta_i M_i = \beta_j M_j$. We can therefore choose M_i for different frequency regions i.e., the spacing of samples in v domain, such that $\beta_i M_i = \beta_j M_j$. The total number of samples, held constant, M_i's are given by

$$\sum_{i=1}^{N} M_i = M_{const} \tag{2.23}$$

$$M_i = \frac{\frac{M_{const}}{\beta_i}}{\sum_{j=1}^{N} \frac{1}{\beta_j}} \tag{2.24}$$

With this choice of M_i's, the non-uniformly spaced samples in v domain are represented as uniformly spaced samples in λ domain. Since the scale is arbitrary in λ domain, we can choose the spacing of samples and origin to some convenient values. Eq. (2.24) shows that the calculation of M_i's depend on β_i's. So, we need to devise a procedure to compute β_i's from the speech data, from which the warping function can be derived easily. So, given the warping parameters (methods of computation of these parameters may be different), we can numerically compute the discrete warping function. Before

examining the method of computation of warping parameters, let us consider the following situation. If there exists a simple linear scaling between two speakers, say $\mathcal{A}$ and $\mathcal{B}$, then $\alpha_{\mathcal{AB}}(f) = \alpha_{\mathcal{AB}}^{\beta(f)} = \alpha_{\mathcal{AB}}$. The scaling factor is only speaker dependent and is independent of frequency. So, $\beta(f) = 1$ or in its discrete form, $\beta_i = 1$, $i = 1, 2, \cdots, N$. The warping function in such a case is given by

$$W(f) = \lambda = \upsilon = \log(f) \tag{2.25}$$

Because of the non-linear scaling between the speakers, the scale factor will be both frequency dependent and speaker dependent. In such cases, $\beta(f)$ models the non-linearity in the scale factor. In other words, the non-linearity in i^{th} frequency region is modelled by β_i. Hence the discrete warping function is given by

$$W_i(f) = \lambda = \frac{\upsilon}{\beta_i} = \frac{\log(f)}{\beta_i}, \ i = 1, 2, \cdots, N \tag{2.26}$$

2.3.2   Band Edge Problem

The discussion in Section 2.3.1 shows that the sampling rate in υ domain changes abruptly at the band edges. Hence, though the sampling is uniform within a given band, it is non-uniform over the whole υ domain. The result is the loss of spectral samples at the band edges. To avoid this loss of spectral samples, we carried the following transition band analysis. Eq. (2.20) shows that the warped spectra in i^{th} frequency region are shifted versions of one another, the shift being frequency dependent, as it is a function of β_i. It is the β_i that determines the spacing of samples in υ domain, whose discontinuity at the band edge results in the loss of spectral samples. One way

to avoid this problem is to make β_p change gradually to β_{p+1}, $p = 1, 2, \cdots, N-1$, over a region of K samples, i.e., we need

$$\frac{\beta_p}{\Delta v_p} = \frac{\beta_p + \Delta\beta}{\Delta v_p + \delta_{p,1}} = \cdots = \frac{\beta_p + k\Delta\beta}{\Delta v_p + \delta_{p,k}} = \cdots = \frac{\beta_p + K\Delta\beta}{\Delta v_p + \delta_{p,K}} = \frac{\beta_{p+1}}{\Delta v_{p+1}} \qquad (2.27)$$

where $\Delta\beta$ is a factor which provides a transition in the values of β across the adjacent regions, K is the number of frequency points over which β_p gradually changes to β_{p+1} defined as $K = \left|\frac{\beta_{p+1} - \beta_p}{\Delta\beta}\right|$, $p = 1, 2, \cdots, N-1$ and $k = 1, 2, \cdots, K$. $\{\delta_{p,k},\ k = 1, 2, \cdots, K\}$ are the factors that are to be computed which provide the gradual change in the sampling intervals across two adjacent frequency regions, thus avoiding the loss of spectral samples at the band edges. Hence, given β_p, β_{p+1} and $\Delta\beta$, the factors $\{\delta_{p,k},\ k = 1, 2, \cdots, K\}$ can be determined from Eq. (2.27). We consider L points to the either side of the band edge over which β_p changes to β_{p+1}, thus amounting to a total of K points, where L is given as

$$L = \frac{K-1}{2} \qquad if\ K\ is\ odd \qquad (2.28)$$

$$= \frac{K}{2} - 1 \qquad if\ K\ is\ even \qquad (2.29)$$

From Eq.(2.21), it is clear that $\Delta v_i \propto \beta_i$. Hence we have the following cases.

1. $\beta_{p+1} < \beta_p$: $\Delta\beta < 0$ and $\{\delta_{p,k},\ k = 1, 2, \cdots, K\}$ forms a decreasing sequence.

2. $\beta_{p+1} = \beta_p$: $\Delta\beta = 0$ and $\{\delta_{p,k} = \delta_{p,k+1},\ k = 1, 2, \cdots, K\}$

3. $\beta_{p+1} > \beta_p$: $\Delta\beta > 0$ and $\{\delta_{p,k},\ k = 1, 2, \cdots, K\}$ forms an increasing sequence.

It is to note that the above analysis to override the band edge effects may not be the optimum method. In our case, β_p varies linearly within the transition band to β_{p+1}. Different variations can be tried out in the transition of β_p within the transition band. Smaller the value of $\Delta\beta$, the transition from β_p to β_{p+1} will be smooth over a large number of points. Once v domain is discretized overriding the band edge effects, the warping function can be computed as explained in Section 2.3.1 using Eq. (2.26) except that the value of β_p in the transition region between p^{th} and $(p+1)^{th}$ frequency regions should be taken as $\beta_i + k\Delta\beta$,where;k=1,2,$\cdots$K.

2.3.3   Experimental Determination of Warping Parameters

The warping parameters $\alpha_{AB}^{(i)}$ and β_i were computed experimentally from the vowel data of PnB and HiL databases. We had chosen N = 5, thus obtaining 5 logarithmically equi-spaced frequency regions. The reason for choosing N to be 5 will be explained later. Table 2.1 shows the frequency regions of interest for PnB and HiL databases. While estimating $\alpha_{AB}^{(i)}$, only two speakers were considered at a time, considering only those pair of formants that lie within the same frequency region. For example, for each pair of speakers, $\mathcal{A}$ and $\mathcal{B}$, we computed the ratio of formants in the ith frequency region as

$$\tau_{AB}^{(i,j,k)} = \frac{F_B^{i,j,k}}{F_A^{i,j,k}} \quad if \quad F_A^{i,j,k}, F_B^{i,j,k} \in [\mathrm{L}_i, \mathrm{U}_i) \tag{2.30}$$

$F_A^{i,j,k}$, $F_B^{i,j,k}$ are the kth formants of the jth vowel of speakers $\mathcal{A}$ and $\mathcal{B}$ respectively and both of them lie in the same i^{th} frequency region. We computed $\tau_{AB}^{(i,j,k)}$ for all pairs

PnB			HiL		
Band (Hz)	β_i	σ_{β_i}	Band (Hz)	β_i	σ_{β_i}
[190,356)	2.13	0.13	[310,524)	1.50	0.03
[356,667)	1.22	0.05	[524,893)	1.55	0.03
[667,1249)	1.51	0.05	[893,1523)	1.46	0.03
[1249,2339)	1.27	0.04	[1523,2598)	1.40	0.02
[2339,4381)	1.00	0.00	[2598,4431)	1.00	0.00

Table 2.1: Average estimates of β_i in 5 logarithmically equi-spaced frequency regions. σ_{β_i} denotes the standard deviation of β_i for i^{th} frequency band. Here, $1 \leq i \leq 5$

of such formants that lie in the ith frequency region and obtained the average scaling factor, $\alpha_{AB}^{(i)}$ as the average of $\tau_{AB}^{(i,j,k)}$ in i^{th} region for a given pair of speakers, $\mathcal{A}$ and $\mathcal{B}$. The estimates of $\alpha_{AB}^{(i)}$ obtained were averaged over to find $\alpha^{(i)}$ representing the frequency dependent scaling factors. β_i's were estimated from the estimates of $\alpha^{(i)}$ as

$$\beta_i = \frac{\log\left(\alpha^{(i)}\right)}{\log\left(\alpha^{(N)}\right)} \qquad 1 \leq i \leq N-1 \qquad (2.31)$$

$$= 1 \qquad i = N \qquad (2.32)$$

Since the higher formants are mostly affected by the length of the pharyngeal cavity, the uniform scaling holds and hence, we assumed $\beta_N = 1$, thus making $\alpha^{(N)}$ to be the ratio of formants in N^{th} frequency region.
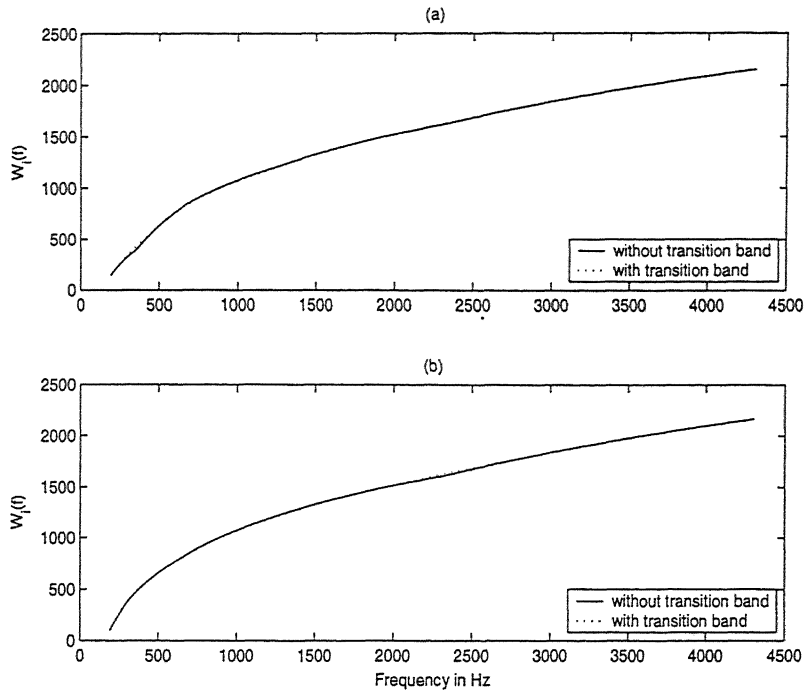
Figure 2.1: Discrete warping function, $W_i(f)$ without transition band and with transition band analysis.

Figure shows the discrete warping function, $W_i(f)$ determined with and without transition band analysis for (a) PnB database and (b) HiL database. Though the curves look very similar, the frequencies at which 'f' axis is sampled are not exactly the same.

2.4   Experiments and Results

The experiments were carried out by overriding the band edge problems to obtain the discrete warping function. Table 2.1 shows the estimates of β_i along with their standard deviations for both PnB and HiL databases, obtained by averaging over all speakers. Figure 2.1 shows the discrete warping function, $W_i(f)$ obtained with and without transition band analysis for PnB and HiL databases. Though, the warping functions look
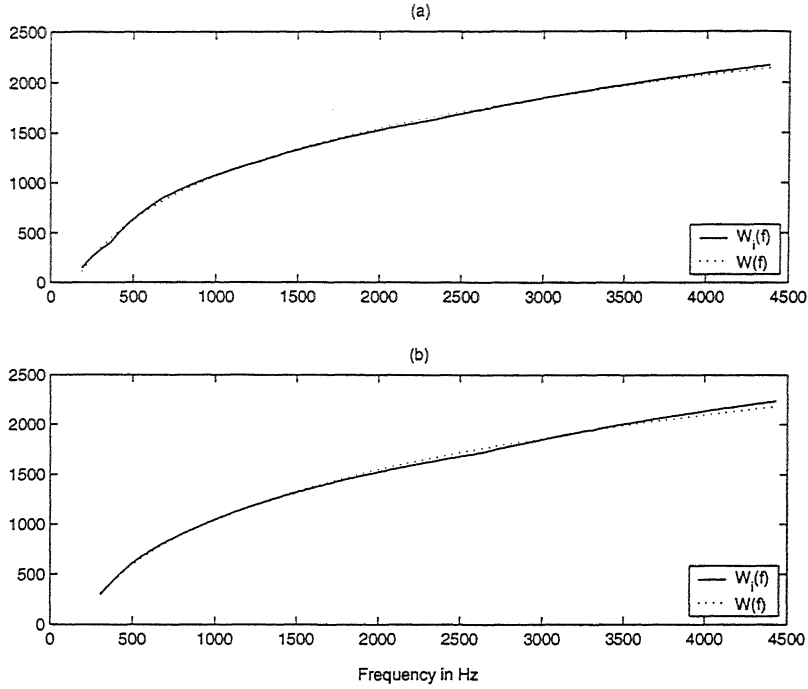
Figure 2.2: Discrete warping function and its closed form approximate, $\mathtt{W}(f)$

Figure shows the discrete warping function, $\mathtt{W}_i(f)$ and its closed form approximate, $\mathtt{W}(f)$ given in Table 2.2 for (a) PnB database and (b) HiL database. The curvefits to $\mathtt{W}_i(f)$ were the best fits obtained from TableCurve2D.

similar, practically, it is important to override the band edge problem. We chose $|\Delta\beta| = 0.0275$ for PnB database and $|\Delta\beta| = 0.010$ for HiL database. Depending on the sign of $(\beta_{p+1} - \beta_p)$, $\Delta\beta$ was chosen to be +ve or -ve for the pth band. The reason for choosing $\mathtt{N}$ to be 5 is to model the non-linearity in a better way. Smaller the value of $\mathtt{N}$, more coarsely will be the modelling of the non-linearity. Larger values of $\mathtt{N}$ results in finer modelling of the non-linearity. But, large values of $\mathtt{N}$ results in less data available for the estimation of $\alpha^{(i)}$ thus questioning its reliability. Hence, a trade-off

Database	$W(f)$
PnB	$-1203.48 + 47.57\,(\log(f))^2$
HiL	$-1323.47 + 49.70\,(\log(f))^2$

Table 2.2: Closed form equations for discrete warping function, $W_i(f)$.
$W(f)$ is the closed form equation for the discrete warping function, $W_i(f)$. The curvefits were obtained by using TableCurve2D package

was to be made between the finer modelling of the non-linearity and the reliability of estimates, resulting in choosing the value of N to be 5. Since the warping function obtained is discrete, we fitted simple curves to it using TableCurve2D to obtain $W(f)$, which is more applicable for continuous spectral patterns. The equations of $W(f)$ for PnB and HiL databases are shown in Table 2.2. Figure 2.2 shows the actual warping function and its closed form approximate, $W(f)$ for PnB and HiL databases. It shows that the curvefits are reliable approximates to their respective originals. Figure 2.3 shows the plot of $W(f)$ for PnB and HiL databases along with mel-warp, log-warp and Stevens & Volkman [1] data points. Mel-warp function defined in Eq. (??) is actually a curve-fit to Stevens & Volkman data points, which were the actual mel frequency data points obtained from psychoacoustic studies. The log-warp function refers to simple linear scaling from the point of view of speaker normalization. The mel scale was derived from psychoacoustic experiments that gave a perceptual measure of pitch. It is a hearing derived scale that relates perceived frequency, and the actual physical frequency. By contrast, the frequency warping function is a speech derived scale that
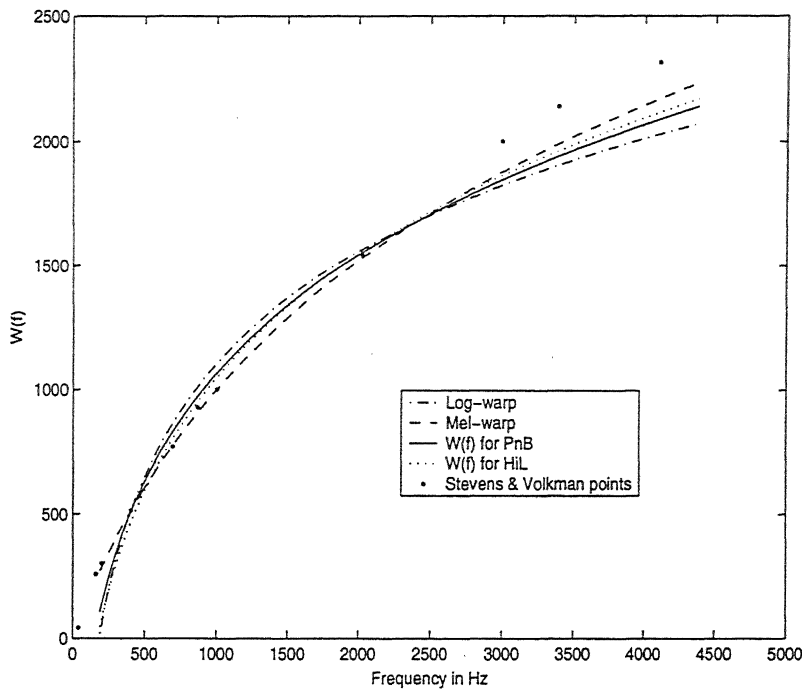
Figure 2.3: Comparison of warping function, W(f), log-warp, mel-warp functions and Stevens & Volkman's actual mel data points.

Figure shows the warping functions for PnB and HiL databases along with log-warp and mel-warp functions. Mel-warp is a function fitted to actual mel data points of Stevens & Volkman. It is interesting to note the similarity of these warping functions, though they are derived from entirely different studies.

maps physical frequency to an alternate domain, λ, such that in the warped domain the speaker dependencies separate out as translation factors. Note the similarity between W(f) and the mel-warp function at frequencies greater than 3500Hz, and between W(f) and log-warp function at frequencies less than 500Hz. In between these frequencies, W(f) lies between mel-warp and log-warp functions, but closer to log-warp than to mel-warp function. This acts as a compromise between the simple linear scaling (based on

speaker normalization) and the mel-scale (based on hearing experiments). This indeed is very interesting that draws some relation between the hearing mechanism and speech production.

2.5 Conclusions

The basic theory of scale invariant transformation was presented. A method for incorporating non-linear scaling in such a paradigm was also discussed. The warping function derived out of the study was compared with log-warp and mel-warp functions which was more log-like at frequencies less than 500Hz and more mel-like at frequencies greater than 3500Hz, and acting as a compromise in between these frequencies.

Chapter 3

Noise Robustness: WOSA Vs. Mel Filter Bank

For all the recognition systems the most common part is the signal processing front end, which converts the speech waveform to some parametric representation for further analysis and processing. The main aim of signal processing front end is to extract such features out of speech which are very robust to variations in speakers and less susceptible to noise. Mainly the most important parametric representation is the short time spectral envelope and therefore spectral analysis methods are considered to be the core of signal processing front end in a speech recognition system. In this chapter we will talk about the noise robustness of weighted overlapped segment averaging (WOSA) and mel filter bank and their mutual comparison in the context of noise robustness for speech recognition.

3.1   Need of Smooth Spectral Envelope

As mentioned earlier the main aim in front end signal processing is to estimate the short time spectral envelope of the phoneme enunciated which actually depends on the vocal tract shape for that particular phoneme enunciated and does not depend on the pitch frequency, which is dependent on the vocal cords vibration. Also we know from source-filter model of speech production that pitch pulses are input to the time varying system representing vocal tract, whose impulse response is smooth spectral envelope in frequency domain, which is different for different phonemes. Therefore our vocal tract is just like a system whose parameters keeps changing with time due to change in the shape of the vocal tract which in turn depends on the phoneme enunciated. Therefore pitch pulses gets convolved with the vocal tract impulse response in time domain and their magnitude fourier tranforms gets multiplied. The effect of this multiplication is sampling of spectral envelope, due to which the spectra of speech for vowels which we get is highly fluctuating. Therefore we can't use this spectra directly for speech recognition rather what we want is the spectral envelope which has been sampled and which is representing the phoneme. Therefore in speech recognition first step is to estimate the smooth spectral envelope, which does not contain any pitch information so that what we get is the actual representation of the phoneme spoken. To remove the pitch periodicities one famous method is homomorphic signal processing of speech spectra i.e. we can remove the the pitch information in cepstral domain by liftering. But in this case in order to get the smooth spectral envelope then we have to get back to

spectral domain and this method becomes computationally expensive. Moreover this processing i.e. liftering of cepstral points is not capable of removing the additive noise. Therefore we need some method which simultaneously removes the pitch periodicities and also helps in reduction of noise from the speech.

In speech recognition the speech is processed frame by frame and frame length is taken to be 20ms as standard but it can be varied like 25ms and 30ms etc. Reason for restricting the frame length to 20ms is that the speech signals are non-stationary and therefore can not be analysed for larger frame lengths. While for 20ms speech signals are supposed to be stationary because the human articulatory system can not respond faster than 20ms.

3.2 WOSA Processing of Speech

WOSA is a well known non-parametric method for spectral estimation. This method was actually popularised by Welch [6]. This method is also used in speech recognition for the estimation of smooth spectra.

3.2.1 WOSA, an Extension of Averaged Periodogram

WOSA processing is actually a variation of the averaged periodogram. In averaged periodogram K independent data records are taken and all of them are the realizations of the same random process and all of them are of same data length L. Let the data are $x_0[n]$, $0 \leq n \leq L-1$; $x_1[n]$, $0 \leq n \leq L-1$; $\cdots x_{K-1}[n]$, $0 \leq n \leq L-1$. Then the

averaged periodogram spectral estimator is defined as

$$\hat{P}_{AVPER}(f) = \frac{1}{K} \sum_{m=0}^{K-1} \hat{P}_{PER}^{(m)}(f) \tag{3.1}$$

where $\hat{P}_{PER}^{(m)}$ is the periodogram for the mth data set .

$$\hat{P}_{PER}^{(m)}(f) = \frac{1}{L} \left| \sum_{n=0}^{L-1} x_m[n] exp(-j2\pi fn) \right|^2 \tag{3.2}$$

If the data blocks are uncorrelated, such an averaging over the data blocks will reduce the variance of the spectral estimate by a factor of K as compared to periodogram estimator and if some correlation exist between the data blocks then variance will not reduce by the same factor. But as periodogram is a biased estimator for correlated signals therefore in averaged periodogram the bias of the estimator will increase because the window length is less as compared to periodogram estimator. But in the special case of white noise, periodogram is an unbiased spectral estimator for any data record length and also the variance of periodogram estimator does not reduce even if we increase the data length. Therefore in case of white noise though averaged periodogram will reduce the variance but there will be no variation in the average value of noise spectral estimate.

As it has been mentioned already that WOSA is just a variation of the averaged periodogram. Its contrasting features are using data window for each data block and the use of overlapped data blocks. Following is the description of WOSA processing of speech signal to get the smooth spectral estimate.
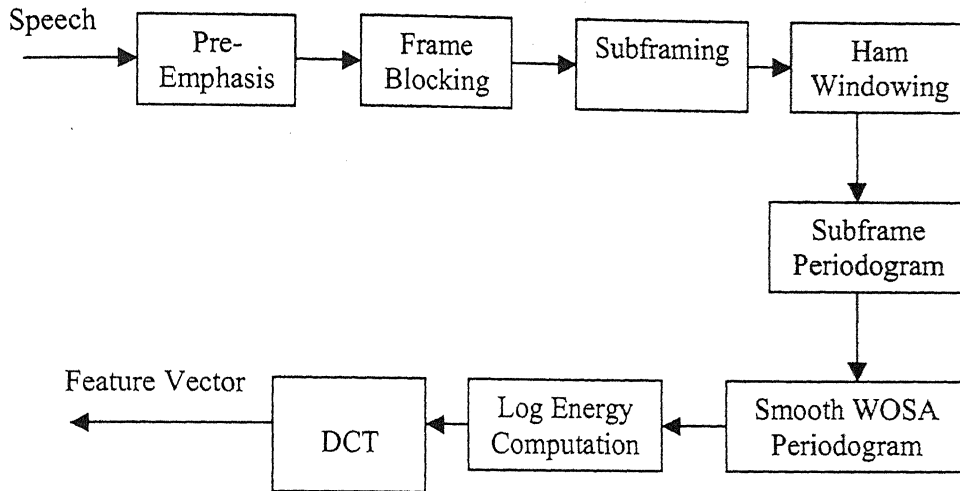
Figure 3.1: Block Diagram of WOSA Processing

3.2.2 Computation of Smooth Spectra and Cepstral Features

Smooth Spectral Estimate : In the context of smooth spectral estimate of speech first of all speech signal is pre-emphasized and then divided into 20ms overlapping frames. Now each of these frames are further divided into overlapped subframes and then each subframe is hamming windowed. Afterward periodogram estimates are found for all the subframes and finally smooth estimate is found by averaging of all the estimates. Hamming window and frame blocking is described in next section. WOSA processing not only helps to remove the pitch periodicities but also helps in reducing the estimator variance.

Cepstral Features: Once smooth spectral estimate is found, it is log compressed and then tranformed into cepstral domain by taking discrete cosine transform(DCT)

of log compressed smooth spectra. Generally only first 13 cepstral coefficients are taken and often the zeroth cepstral coefficient is replaced by normalised energy of the speech frame. Normalised energy is found by dividing the energy of that speech frame with respect to maximum energy over all frames in the utterance. Now this ratio is log compressed and the least value is fixed at -50 because after taking log the values will be ≤ 0. After this one is added to this so as to make the highest energy frame one and then this value is scaled by 0.1 . And finally these thirteen cepstral coefficients are used as the feature vector. Velocity and acceleration features of these cepstral features are also used on the fly during speech recognition.

3.2.3 Noise Robustness and Pitch Removal Quality of WOSA

Actually in a frame of speech of 20ms generally three or four pitch pulses are found for male speaker and for female speaker they will be further more. Therefore by subframing the speech frame the number of pitch pulses will be around one which will not effectively sample the spectral envelope. As mentioned earlier in averaged periodogram that it reduces the variance of estimator for uncorrelated data blocks. Now as PSD of any noisy speech frame is addition of PSD of clean speech (which is correlated for subframes of a frame) and PSD of white noise (which is uncorrelated for subframes of a frame) i.e. if s(n), w(n) and x(n) represents speech signal, noise signal and noisy speech signal respectively then

$$x(n) = s(n) + w(n) \tag{3.3}$$

The short time power spectral density relation is given by

$$P_x(f) = P_s(f) + P_w(f) \tag{3.4}$$

where P represents the PSD of respective subscript.

Therefore by averaging the subframe spectral estimates of noisy speech, the variance of the noise PSD is reduced by number of subframes, while the variance of the PSD of clean speech is reduced by a very less amount. This is just like suppressing the noise and relatively boosting the speech signal. But at the same time this method has the problem of increase in bias of the estimator due to small subframe data length as there is correlated speech signal in noisy speech for which periodogram is a biased estimator.

Therefore by varying the frame length, subframe length and overlap we can check that which combination suits our requirement and we have to compromise between variance and bias, well known as bias-variance tradeoff.

3.2.4 Filter Bank Interpretation of WOSA

The another way of looking at the WOSA is in fourier domain. As we know that multiplication of speech by a window in time domain is equivalent to convolution of speech with window in frequency domain. Therefore fourier transform of windowed speech at some frequency f is the sum of weighted speech energy from a filter at center frequency f and this filter is actually fourier transform of window e.g. for rectangular

window it is sync. For a N point FFT the number of filter banks are equal to the spectral points estimated. In this chapter the effects of WOSA in reducing noise have been studied for various frame lengths, subframe lengths and subframe overlaps over various SNRs.

3.3 MFCC Parameter Computation

In this section we will talk about the mel filter cepstral coefficients (MFCC) feature computation for speech recognition.

3.3.1 Introduction to MFCC

As the name suggest mel filter cepstral coefficients (MFCC) is based on filterbank theory i.e. it consist of N bandpass filters covering the frequency range of interest in the signal. The output of each filter bank is the energy in the spectra covered by them around their central frequency. These filters are constant Q filters i.e. the bandwidth of the filters increases with increase in the center frequency of filters. Moreover these filters are overlapping. The typical structure of the 15 channel MEL filterbanks has been shown in Figure 3.2.

Mainly there are two types of filter bank used for speech recognition, namely uniform filter bank and non-uniform filter bank. In uniform filter bank all filters are spaced uniformly over the frequency range, while in non-uniform filter bank filters are spaced non-uniformly according to some criterion. One commonly used criterion is
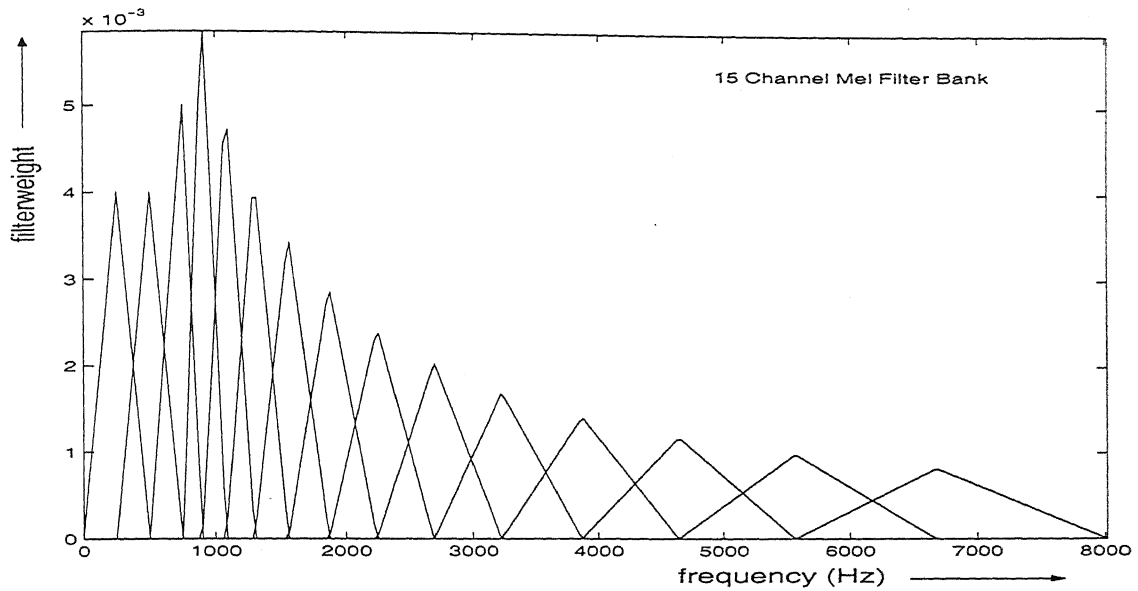
Figure 3.2: 15 Channel MEL Filter Bank

to space the filters uniformly on logarithmic frequency scale which is often justified from a human auditory perception point of view. In MFCC, the MEL scale is used to place the filters. The relation between linear frequency and mel frequency is given by Shaughnessy formula as

$$\nu = 2595 log_{10}(1 + \frac{f}{700})$$

(3.5)

Where ν is the frequency in the warped domain and f is the linear frequency.

3.3.2 Basic Steps in MFCC Feature Computation

Block diagram of MFCC feature computation has been shown in figure 3.3.

1. *Pre-emphasis, Frame Blocking and Windowing:* First of all the speech signal is pre-emphasized so as to boost the higher order formants. The pre-emphasis filter is a
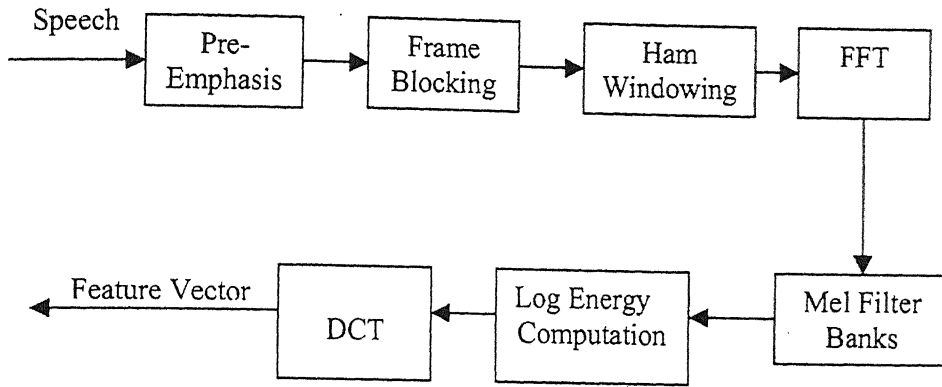
Speech → Pre-Emphasis → Frame Blocking → Ham Windowing → FFT

Feature Vector ← DCT ← Log Energy Computation ← Mel Filter Banks

Figure 3.3: MFCC Feature Computation

first order FIR filter as given below

$$H(z) = 1 - \alpha z^{-1}, \qquad 0.9 \leq \alpha \leq 1.0 \qquad (3.6)$$

After this pre-emphasised speech signal is blocked into frames of length N and adjacent frames being separated by M samples and $M \leq N$ which leads to the overlapped frames. The next step is to hamming window each frame so as to reduce the signal discontinuities at the beginning and the end by using the tapered hamming window. Hamming window w(n) is given as

$$w(n) = 0.54 - 0.46 cos(\frac{2\pi n}{N-1}), \qquad 0 \leq n \leq N-1 \qquad (3.7)$$

Computation of Energy Vector: Now each windowed frame is transformed into frequency domain and then energies are found for each filter bank, which are further log compressed so as to reduce the dynamic range and apparently to further smooth out the spectrum more.

log compressed so as to reduce the dynamic range and apparently to further smooth out the spectrum more.

Computing Cepstral Coefficients: Finally the log spectral energies are transformed into cepstral domain by taking their discrete cosine transform. This has the effect of decorrelating the log spectral energies and also reduces the dimensionality of the feature vector. Here in MFCC also zeroth coefficient is replaced by normalised energy as described previously. In MFCC also pitch information is removed by taking the weighted filter bank energies as the output, which actually is a weighted average of the speech signal around center frequency and over a spread equal to bandwidth of the filter bank. MFCC also helps in the removal of pitch information and reduction in noise estimate variance, as here also same averaging is carried out over the spread of filter bank for all the channels.

3.3.3 Differences in Filter Bank Structure

Though WOSA and Mel filter bank both have filter bank interpretation yet, they differ in the filter bank structures. As discussed earlier that windowing in time domain is convolution in frequency domain and this way WOSA can be interpreted as filter bank in the frequency domain. In case of WOSA, filters are spaced uniformly on linear frequency scale while in Mel filter bank, filters are spaced uniformly on Mel scale. WOSA filter bank have constant band width, while MFCC filter bank have constant Q. The rationale behind constant Q filter bank is that at higher frequencies the ear's

frequency resolution is poor.

3.4 Experimental details

In the present study, the capability of WOSA in extracting features robust to the white noise has been studied for various frame lengths, subframe lengths, subframe overlaps and over various SNRs. The recognition accuracies were considered as the measure. The results were compared with 26 channel Mel filter bank. Both of these methods were tested for vowel recognition and digit recognition. For recognition experiments the models were based on hidden Markov model (HMM) and HTK (HMM ToolKit) was used for simulation of speech recognizer. For vowel recognition TIMIT database was used and experiments were carried out for mixed gender. While for digit recognition Numbers corpus was used. The models were prepared from clean speech and tested on clean and noisy speech at various SNRs, where noise is AWGN. In vowel recognizer 12 vowels were modeled and in digit recognition 20 phonemes were modeled. Timit database was sampled at 16KHz and Number corpus database was sampled at 8KHz, therefore 20ms is equal to 320 samples and 160 samples respectively. In WOSA the experiments were carried out for various subframe lengths, subframe overlaps and SNRs. Frame lengths were taken to be 20ms and 25ms with 50% overlap. Moreover in WOSA log warping was used and number of spectral points were 96. In MFCC the channels were fixed to be 26 in vowel recognizer and 21 in digit recognizer with filters spaced uniformly on mel scale based on Shaughnessy formula given in equation 3.5.

SNR(dB)	1st	2nd	3rd
clean	53.41 (160-0)	53.00 (MFCC)	52.90 (128-64)
36	53.81 (160-0)	53.10 (MFCC)	52.95 (160-120)
24	50.86 (160-140)	50.71 (160-120)	50.46 (MFCC)
6	36.47 (160-140)	36.01 (160-120)	35.91 (128-64)
-6	24.06 (160-140)	23.65 (160-120)	21.97 (128-64)

Table 3.1: Vowel Recognition Accuracies for 20ms frame length

Table enteries are given as: Percent Accuracy (Subframe length - Overlap)

The different SNRs for all the testing experiments were fixed to be 36 dB, 24 dB, 6 dB, -6 dB and clean speech.

The various subframe lengths in WOSA, in vowel recognition, for 20ms (320) frame length, were 6ms (96), 8ms (128) and 10ms (160). while for 25ms (400) frame length, subframe lengths were 6.25ms (100), 9.375ms (150) and 12.5ms (200). In digit recognition, the experiments were conducted for 20ms (160) frame length and various subframe lengths used were 8ms (64), 10ms (80) and 16ms (128). Within brackets are the number of samples for the corresponding frame length in ms.

3.5   Results and Conclusions

Results The results for vowel recognizer are shown in Table 3.1 and 3.2. Results of digit recognizer are shown in Table 3.3. In result tables only best three signal processings in terms of recognition accuracies for different SNRs have been shown.

Conclusions: As in most of the cases WOSA performed better than Mel filter bank

SNR(dB)	1st	2nd	3rd
clean	55.21 (200-0)	53.35 (200-100)	53.14 (200-150)
36	53.71 (200-0)	53.56 (200-150)	53.45 (200-100)
24	51.29 (200-0)	50.10 (150-25)	50.05 (150-68)
6	34.79 (150-100)	34.38 (150-25)	34.33 (100-80)
-6	24.79 (200-150)	23.97 (200-100)	22.89 (200-00)

Table 3.2: Vowel Recognition Accuracies for 25ms frame length

Table enteries are given as: Percent Accuracy (Subframe length - Overlap)

SNR(dB)	1st	2nd	3rd
clean	94.41 (80-0)	94.40 (80-40)	94.35 (64-32)
36	94.17 (80-70)	94.11 (80-60)	94.03 (64-45)
24	92 (80-70)	91.98 (64-32)	91.95 (64-45)
6	60.22 (80-70)	60.05 (128-124)	60.02 (80-60)
-6	23.38 (80-70)	23.13 (80-60)	22.28 (64-45)

Table 3.3: Digit Recognition Accuracies for 20ms frame length

Table enteries are given as: Percent Accuracy (Subframe length - Overlap)

therefore we will discuss only about WOSA. Following conclusions can be drawn from the result tables.

1. From vowel recognizer results it can be seen that in most of the SNRs, 25ms frame length has shown slightly better performance than 20ms frame length. But the frame length may not be increased further because in that case non-stationarity of speech may dominate and performance may start degrading.

2. In all the three tables it can be seen that recognition accuracy is better for 50%

subframe length as compared to others. This is due to the reason that if we decrease the subframe length further, then more bias will creep into the estimate which will degrade the performance.

3. From the tables it can be seen that increased overlap perform better at low SNRs. The increased overlap in turn increases the number of subframes being averaged. This behaviour can be explained with the reason that as noise increases the estimator variance also increases and to reduce that variance i.e. to make the spectral estimate smooth, more averaging is required at lower SNRs. But if we further increase the overlap then the performance starts degrading because more averaging will carry out oversmoothing of spectra.

4. Even though WOSA shows relatively better performance yet, there is a large degradation in the performance as the SNR decreases. This is due to the reason that though WOSA does reduce the estimator variance yet, it has no effect on the average psd of the white noise part in the noisy speech psd . Therefore, though WOSA is reducing the estimator variance, it is not capable of reducing the average noise i.e. mean of noise estimate. Moreover, shortening of window introduces more bias due to the presence of correalated speech signal and the large overlap used to carry out the oversmoothing of the spectra. Similar observations can also be made about Mel Filter Bank feature computation method.

Chapter 4

Cross Correlation Based Noise Reduction

Though there are various types of noises [12] which can effect the overall recognition performance yet, background noise is a major problem in speech recognition. We have seen in the previous chapter that the effect of noise on speech recognition is very adverse. Both WOSA and Mel filter bank do not extract such features that are robust to noise. Although many noise reduction techniques have been proposed yet, the performance of existing speech recognition systems still degrades rapidly in the presence of noise.

4.1 Spectral Subtraction Techniques for Noise Reduction

Some of the early works done in the area of noise reduction were done by Widrow [3] in 1975 and Boll [4] in 1979 where they focussed on the idea of speech enhancement. These methods are also known as signal enhancement preprocessor. Later on these methods were also used for speech recognition. Athough many other methods have been proposed for noise reduction, in the current study we have focussed on spectral subtraction method only .

4.1.1 Boll's Method of Noise Estimation

The Boll's method is based on the criterion that the noise is stationary and there are pauses during the utterance where the speech energy is very negligible. These speech pauses are then used for the estimation of noise magnitude spectra. In his method, for stationary case, noise is estimated from first few ms (typically 100ms), supposing that there is no speech activity during that part. If the noise is slowly varying then his method requires the Voice Active Detector (VAD) so that noise can be updated.

Widrow's method is about adaptive noise cancellation. He makes use of the two microphones, one of which is used for receiving speech plus noise and another one is used for picking up only noise. Then this noise is passed through an adaptive filter and finally the noise estimate is subtracted from noisy speech.

All of the methods of spectral subtraction are based on the assumption that the

background noise is additive and uncorrelated to the speech signal. Let S(n), W(n) and X(n) represents the speech signal, noise signal and noisy speech signal respectively.

$$X(n) = S(n) + W(n) \tag{4.1}$$

and short time power spectral density relation is thus given as

$$P_x(f, i) = P_s(f, i) + P_w(f, i) \tag{4.2}$$

Where $f \in [0, \text{N-1}]$ is a discrete variable enumerating the frequency bins and i is the time block index. The short time power spectral density is estimated by using the periodogram as described in last chapter and given below for more clarity.

$$\hat{P}_x(f, i) = \frac{1}{N} |\mathcal{F}[X_N(i)]|^2 \tag{4.3}$$

Where $X_N(i)$ is a noisy speech vector containing the ith block of N data samples and $\mathcal{F}$ is FFT operation.

Therefore all of these methods try to compute smooth estimate of the noise from some portions of utterance, where speech activity is not present. The rationale behind finding the smooth estimate of noise is to reduce the error in mean square sense. To find the smooth spectral estimate of noise from pause portion,that portion is divided into frames of some standard length (20 ms in GSM) and then periodogram estimate is found for each frame. Finally estimate is averaged out over all the frames to get the smooth estimate.

4.1.2 A Different Approach of Noise Estimation

As mentioned in previous section that smooth estimate of noise spectra minimizes the mean square error, because when we don't have access to a particular realization of noise from AWGN process then best thing which we can do is to estimate the mean of the noise spectra from somewhere else assuming that mean does not change (WSS) and to subtract that estimate from the noisy speech PSD. If we look at the typical noise PSD periodogram estimate for a frame of noise only in figure 4.1, one can easily see that there will be residual noise even after spectral subtraction. If somehow we could estimate this highly varying noise PSD for each frame then this would have been best way. But this is not possible because as what we get is the speech plus noise. Therefore we think in terms of reducing error, assuming some error criterion. In this case used error criterion is mean square error. In the proposed cross correlation based method of noise estimation the basic assumption is the same that noise is uncorrelated with the speech and it is additive. To estimate the noise this method makes use of two adjacent frames of noisy speech. Let $X_1(n)$ and $X_2(n)$ are two adjacent frames of noisy speech then

$$X_1(n) = S_1(n) + W_1(n) \tag{4.4}$$

$$X_2(n) = S_2(n) + W_2(n) \tag{4.5}$$

Where S and W are representing clean speech and white noise respectively. Now assuming uncorrelatedness of noise with speech, the auto-correlation of each frame and
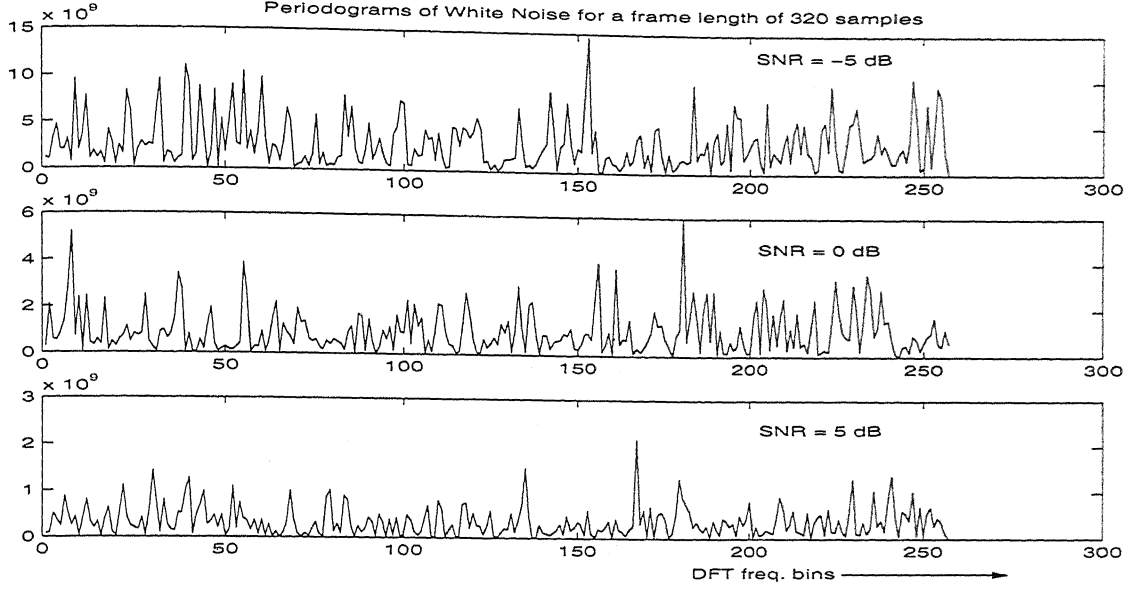
Figure 4.1: PSD of Noise at Various SNRs

cross correlation of two adjacent frames can be written as below

$$X_1(n) \star X_1(n) = S_1(n) \star S_1(n) + W_1(n) \star W_1(n) \tag{4.6}$$

$$X_2(n) \star X_2(n) = S_2(n) \star S_2(n) + W_2(n) \star W_2(n) \tag{4.7}$$

$$X_1(n) \star X_2(n) = S_1(n) \star S_2(n) \tag{4.8}$$

Where $\star$ is representing correlation operation. In these equations noise and speech have been assumed to be uncorrelated and therefore all cross correlation term of noise and speech have been neglected.

These equations in PSD domain can be written as

$$\hat{P}_{X1,X1}(f) = |S_1(f)|^2 + |W_1(f)|^2 \tag{4.9}$$

$$\hat{P}_{X2,X2}(f) = |S_2(f)|^2 + |W_2(f)|^2 \tag{4.10}$$

$$\hat{P}_{X1,X2}(f) = S_1(f)S_2^*(f) \tag{4.11}$$

The Eq.4.11 represents the cross PSD of two adjacent frames. To find out the noise estimate for a frame, we define $W_{est}(f)$ using above Equations in the following way

$$W_{est}(f) = 0.5 * (\hat{P}_{X1,X1}(f) + \hat{P}_{X2,X2}(f) - 2 * |\hat{P}_{X1,X2}(f)|). \tag{4.12}$$

$$W_{est}(f) = 0.5 * (|S_1(f)|^2 + |S_2(f)|^2 - 2 * |S_1(f)||S_2(f)| + (|W_1(f)|^2 + |W_2(f)|^2)) \tag{4.13}$$

Which can be further simplified as

$$W_{est}(f) = 0.5 * ((|S_1(f)| - |S_2(f)|)^2) + 0.5 * (|W_1(f)|^2 + |W_2(f)|^2) \tag{4.14}$$

Equation 4.14 contains two terms. First term is half of the square of the difference of magnitude fourier transform of clean speech in two adjacent frames and second term is the half of the sum of noise PSD estimates of same two adjacent frames. The difference term is small because there is good correlation between two adjacent clean speech frames, but this term is not negligible. Actually this first term is supposed to be smaller due to two reasons, the first is that the difference of magnitude fourier transform of two adjacent clean speech frames is assumed to be small due to correlation, second it is relatively smaller at lower SNRs as compared to the second term. But at higher SNRs (e.g 15 dB, 20 dB) this difference term becomes significant as compared to second term. Therefore at higher SNR this term creates problem. Now if we can somehow get rid of this term then second term will become our local estimate of noise. Local because this estimate is only valid for these two frames. The effect of this difference term is highly dominating in the First formant region, specially for vowels. This thing

$$\hat{P}_{X1,X2}(f) = S_1(f)S_2^*(f) \tag{4.11}$$

The Eq.4.11 represents the cross PSD of two adjacent frames. To find out the noise estimate for a frame, we define $W_{est}(f)$ using above Equations in the following way

$$W_{est}(f) = 0.5 * \left(\hat{P}_{X1,X1}(f) + \hat{P}_{X2,X2}(f) - 2 * |\hat{P}_{X1,X2}(f)|\right). \tag{4.12}$$

$$W_{est}(f) = 0.5 * \left(|S_1(f)|^2 + |S_2(f)|^2 - 2 * |S_1(f)||S_2(f)| + (|W_1(f)|^2 + |W_2(f)|^2)\right) \tag{4.13}$$

Which can be further simplified as

$$W_{est}(f) = 0.5 * \left((|S_1(f)| - |S_2(f)|)^2\right) + 0.5 * \left(|W_1(f)|^2 + |W_2(f)|^2\right) \tag{4.14}$$

Equation 4.14 contains two terms. First term is half of the square of the difference of magnitude fourier transform of clean speech in two adjacent frames and second term is the half of the sum of noise PSD estimates of same two adjacent frames. The difference term is small because there is good correlation between two adjacent clean speech frames, but this term is not negligible. Actually this first term is supposed to be smaller due to two reasons, the first is that the difference of magnitude fourier transform of two adjacent clean speech frames is assumed to be small due to correlation, second it is relatively smaller at lower SNRs as compared to the second term. But at higher SNRs (e.g 15 dB, 20 dB) this difference term becomes significant as compared to second term. Therefore at higher SNR this term creates problem. Now if we can somehow get rid of this term then second term will become our local estimate of noise. Local because this estimate is only valid for these two frames. The effect of this difference term is highly dominating in the First formant region, specially for vowels. This thing

is justified as the difference term will be significant in a region where speech energy is mainly clustered i.e. first formant region, for many vowels.

Therefore if we can remove this difference term mainly from this region then we will be left with the second part of equation 4.14, which is our noise estimate. If we can filter out the first term then second term is our noise estimate. We can compare this estimate with smooth estimate, in terms of mean square error, to justify it theoretically.

As we know that smooth estimates tries to estimate mean of the noise psd in the speech signal i.e. expected value of the estimator. Also the periodogram estimator of white noise at each spectral point is having independent identical distribution (iid). Therefore the MSE of noise estimates over all the spectral points with respect to original noise psd embedded in the noisy speech signal has been given below. Let us assume that σ_{est}^2 is the variance of the estimator and μ_{est} is the mean of the estimator i.e. desired smooth estimate. If this μ_{est} is subtracted (assuming that it has been truly estimated from pauses) from the noise in a frame then mean square error will be

$$
\begin{aligned}
MSE &= \mathcal{E}[(\hat{P}_w(f) - \mu_{est})^2] \\
&= \sigma_{est}^2
\end{aligned}
\tag{4.15}
$$

Where $\mathcal{E}$ is expectation operator and $\hat{P}_w(f)$ is noise PSD part in the PSD estimate of noisy speech in equation 4.2. Therefore MSE in that case is equal to the variance of the estimator. While in the proposed case we know that noise estimate is equal to the mean of the noise in two adjacent frames and one of them is the noise to be

estimated.There error in this case is equal to

$$
\begin{aligned}
MSE &= \mathcal{E}[(\hat{P}_{W1}(f) - 0.5 * (\hat{P}_{W1}(f) + \hat{P}_{W2}(f)))^2] \\
&= \mathcal{E}[(0.5 * (\hat{P}_{W1}(f) - \hat{P}_{W2}(f)))^2] \\
&= \frac{1}{2}\sigma^2_{est}
\end{aligned}
\qquad (4.16)
$$

Therefore in this case under the assumption that we somehow filter out the first term in the equation 4.14, mean square error reduces to half of that in smooth estimate case.

4.2 Experimental Details

As mentioned in the previous section, that one main problem in estimating noise in this way is the difference term and also it has been shown that this difference is significant in the lower frequency domain. Therefore one way to reduce the effect of this term is to suppress this term in lower frequency domain. For that purpose we can carry out pre-filtering in frequency domain so as to reduce the effect of difference term.

Pre-Filters: As the relative effect of the difference term is not same at all the SNRs, therefore different filters are required at different SNRs. But to find out the SNR is difficult as we don't have any prior knowledge of noise. But if we look at the figure 4.2, the Inverse Fourier Transform of estimated noise in equation 4.14, then it is observed that the ratio (shown in figure 4.2) of zero lag peak to next peak in the neighbourhood of peak at zero lag is increasing with the decrease in the SNR. Though
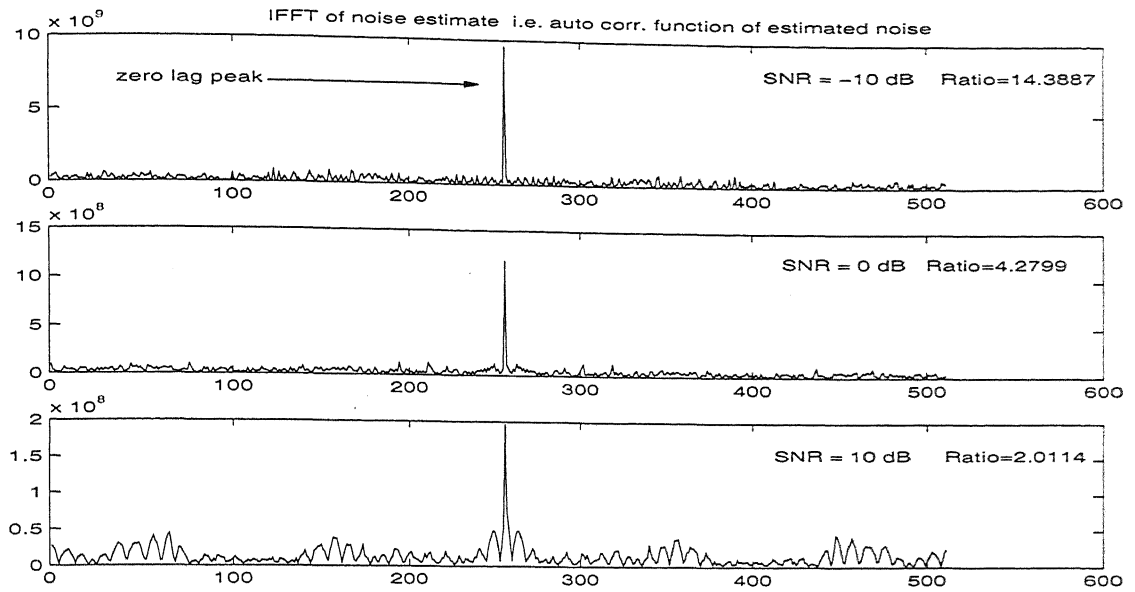
Figure 4.2: IFFT of Noise Estimate

this experiment was not carried out in the present study but this ratio can be used

for the range estimation of SNR i.e. with the help of this ratio we can tell about the

range of SNR. But this ratio can't be used for point estimation of SNR because this

ratio has also some variance.

To find out the filters we found the ratio of noise psd to be estimated $\hat{P}_{W1}(f)$ (as

during experimentation we had access to added noise) and estimated noise psd W_{est},

at each frequency bin for many files and frames. These ratios for a particular SNR

are stored and analysed on histogram basis. Any ratio value above 1000 was left as

an outlier. Finally ratios corresponding to peaks of the histogram were chosen as filter

coefficient in frequency domain, separately for each frequency bin and for each SNR.

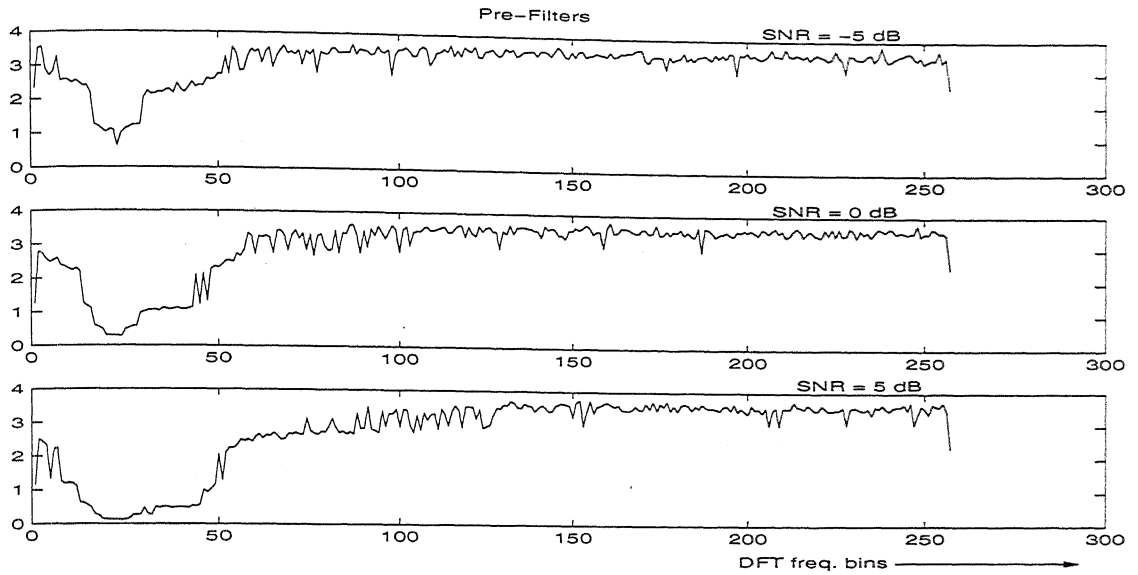These filters were fixed for the rest of the experiments. Few of the filters are shown

Figure 4.3: Pre-Filters

in figure 4.3. Once noise is estimated we can use it for speech enhancement or speech recognition. In this study we have used it for vowel speech recognition and TIMIT DR2 data base has been used. Basics steps to estimate noise and carry out speech recognition has been shown in figure 4.4 and described below.

Pre-emphasis, Frame-Blocking and Windowing: This is the same as described in previous chapter.

Estimation of Noise PSD and Enhanced Speech: Then next step is to estimate the PSD's of two adjacent frames and their Cross PSD. Periodogram estimator has been used for this purpose. Cross PSD estimation of a frame has been done with previous frame. Noise is estimated using equation 4.14. Now this initial noise estimate is filtered with the help of pre-filters found above. This noise estimate is subtracted from the
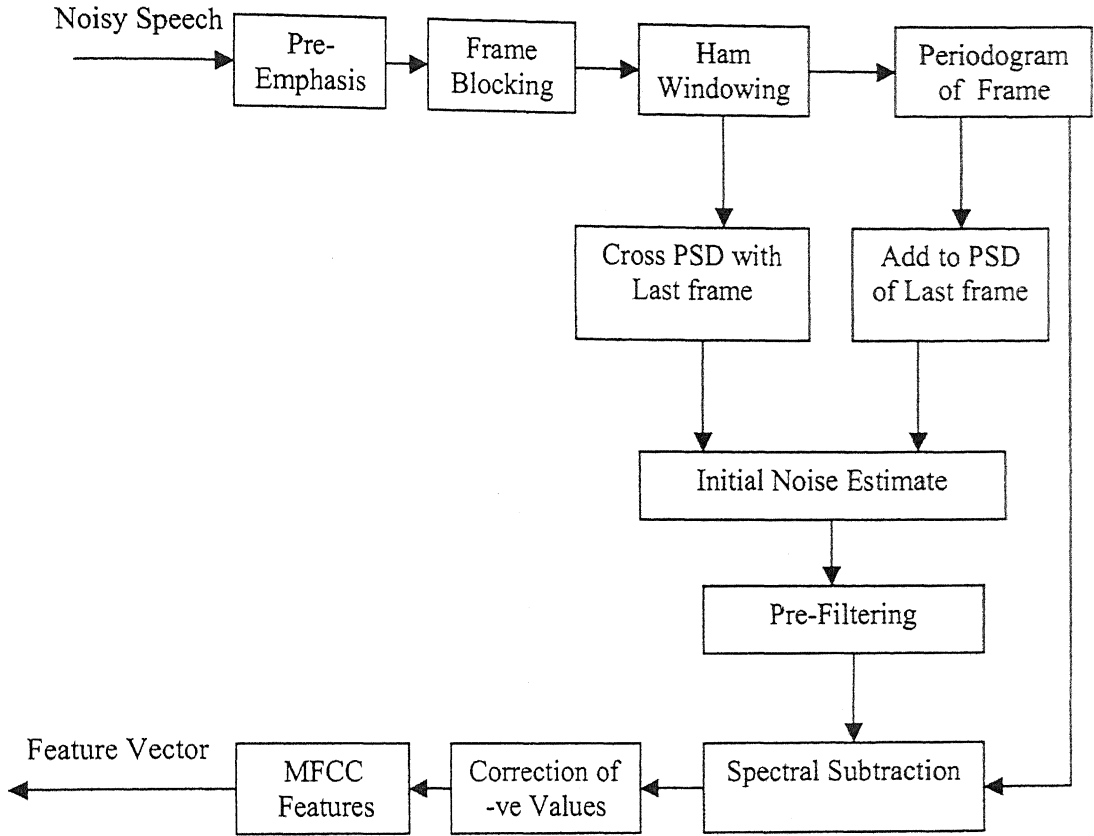
Figure 4.4: Block Diagram of Noise Estimation and Feature Computation

PSD estimate of noisy speech frame for which this noise estimate was found.

Correction For negative amplitude: After subtraction of noise estimate, there may be negative amplitudes in the enhanced PSD. To correct them the whole frequency region is divided into five frequency bands and if there is a negative amplitude in any band, that is replaced by β times the minima of noisy speech PSD in that band. Where β is less than one.

MFCC features: Once we have found the estimate of clean speech, then we need to

find the cepstral coefficients for recognition. For this experiment we used the MFCC features. MFCC features were computed in the same way as explained in previous chapter.

Following this procedure, noisy speech was enhanced and features were computed for that. HMM models were trained for 12 vowels with features drawn out of clean speech and after that tested on clean speech, noisy speech and enhanced speech at various SNRs. Training and testing data sets were totally different. The results and conclusions of the experiments conducted have been discussed in next chapter.

Chapter 5

Results and Conclusions of Proposed Noise Compensation Method

5.1 Results

We carried out the experiments on a vowel recognizer. Noisy speech, enhanced speech and clean speech has been shown in figure 5.1, 5.2, 5.3 and 5.4 for few frames of speech at various SNRs. Although from the figures it may seem that there is no information in the higher frequency region, that is actually not true. This is because even the spectral amplitudes in high frequency region are small, they are important in characterizing fricatives and consonants. In our experiments, we have developed statistical models for 12 vowels using HMM ToolKit (HTK). Our proposed method was tested only for low SNRs, -10 dB to 10 dB. Models were prepared from clean speech and tested on clean speech, noisy speech and enhanced speech for the purpose of comparison.
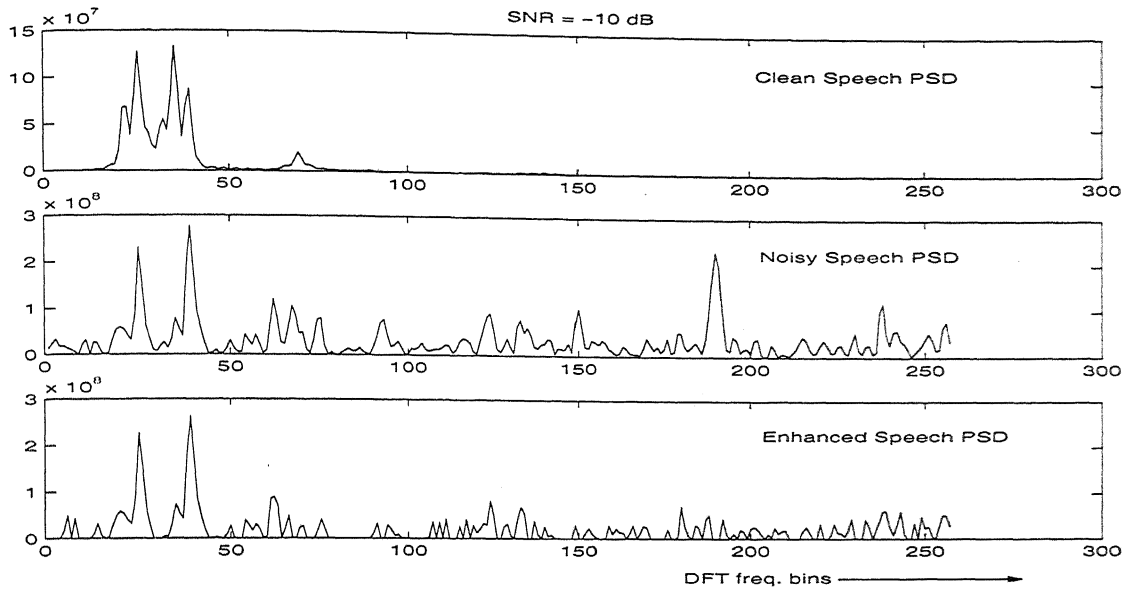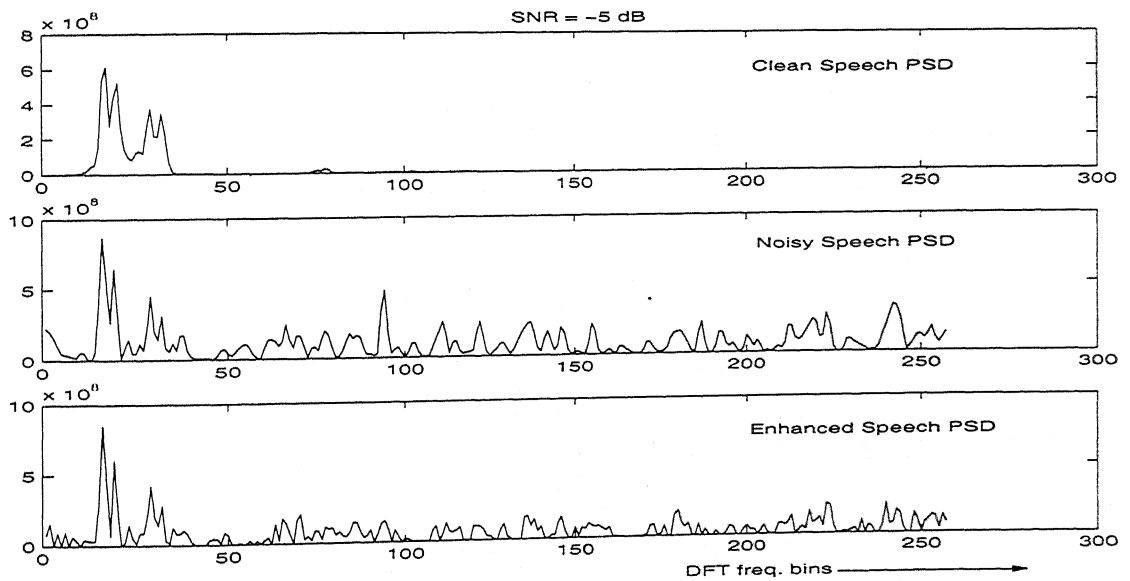
Figure 5.1: Enhanced Speech frame at -10 dB

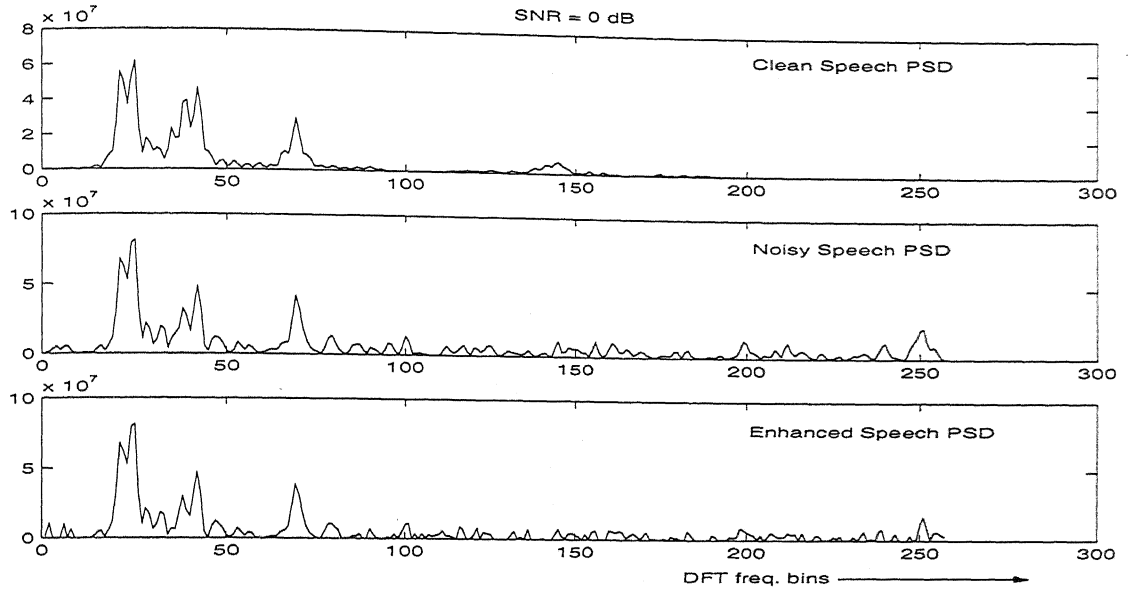Figure 5.2: Enhanced Speech frame at -5 dB
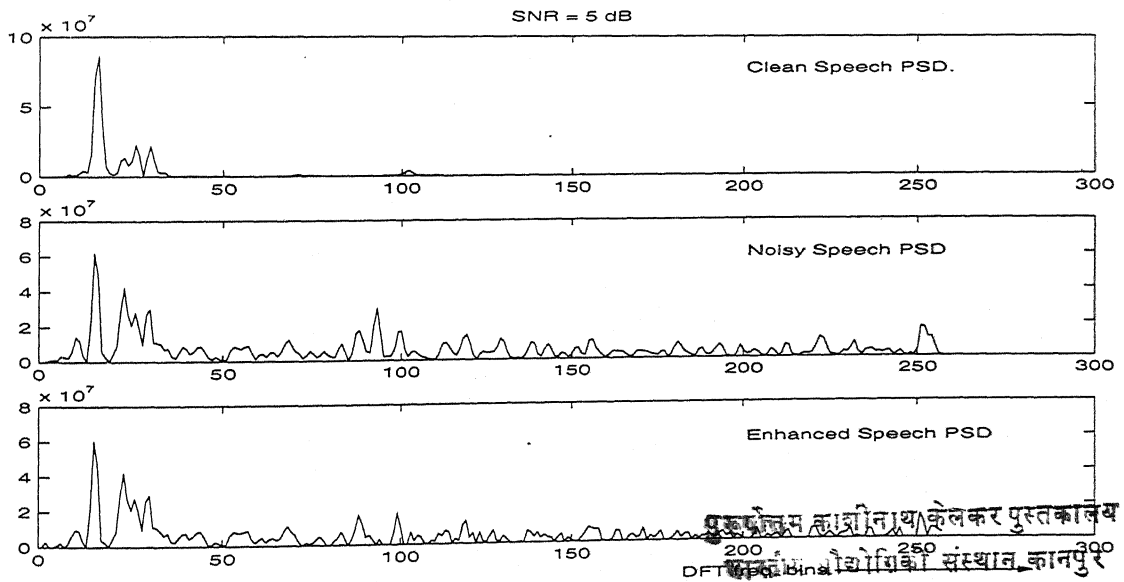
Figure 5.3: Enhanced Speech frame at 0 dB

Figure 5.4: Enhanced Speech frame at 5 dB

SNR(dB)	Noisy SP.	Enhanced Sp.
10	34.74	34.29
5	20.40	21.88
0	13.75	15.59
-5	13.16	13.30
-10	13.08	13.08

Table 5.1: % Recognition Accuracies, Without Pre-emphasis
Recognition Accuracy for Clean Speech is 60.31%

SNR(dB)	Noisy SP.	Enhanced Sp.
10	50.26	49.22
5	38.29	36.73
0	20.84	25.50
-5	13.60	15.30
-10	13.08	13.30

Table 5.2: % Recognition Accuracies, With Pre-emphasis
Recognition Accuracy for Clean Speech is 59.42%

The training and test data sets were totally different. First, vowel recognition was carried out without pre-emphasis of speech. The recognition accuracies, without pre-emphasis, has been shown in Table 5.1.

Second, vowel recognition was carried out with pre-emphasis of speech. The recognition accuracies for this experiment has been shown in Table 5.2

5.2   Conclusions

1. The recognition results with pre-emphasis are better than those without pre-emphasis. This can be justified as pre emphasis will boost the signal in higher frequency domain and therefore has boosted the recognition accuracies.

2. From recognition results it is clear that this new cross correlation based noise estimation method has not shown much better performance over the baseline i.e. recognition accuracies for noisy speech. One reason of this non-improvement is the effect of cross correlation terms, which were assumed to be zero on the basis of uncorrelatedness of noise. These terms e.g. speech and noise correlation and cross correlation of noise in two frames are actually not zero and terms becomes significant at very low SNRs like -10 dB and -5 dB.

3. Another possible reason for this non-improvement is the difference term in the noise estimate equation 4.14. This term also starts dominating as the SNR goes high. Though we tried to filter out this term with pre-filters yet, at higher SNRs it could not improve the results for the above mentioned reason.

4. We can observe from the results that the accuracy at 10 dB and 5 dB is slightly less than the baseline. This can be justified for reason given in point 3. At 0 dB

the result has improved by approximately 5% in table 5.2. Similarily at -5 dB it has improved by a little amount of 1.7%, and at -10 dB it has shown negligible improvement.

5. The reason for 5% improvement at 0 dB is that at this SNR both of these problems mentioned previously were not much effective as difference term is effective at higher SNRs and cross correlation terms are significant at very low SNRs.

6. The difference term in equation 4.14 can be filtered out with more better pre-filters. Actually the job of pre-filters is twofold, one it tries to filter out the difference term and secondly it modifies the noise estimate disturbed by cross correlation terms. Therefore with the proper choice of pre-filters. This method can be made to work at least for the SNR range of -5 dB to 10 dB.

5.3   Future work

1. As mentioned in the conclusions, the choice of pre-filters, for filtering out difference term in equation 4.14 and for reducing the effect of cross correlation terms, is a major problem. Therefore more studies need to be carried out to design these pre-filters.

2. This method was tested only on vowel recognizer and needs to be tested on consonants. Further, studies can also be made on the effect of the proposed method on a digit recognizer

3. As mentioned in 4th chapter, the ratio of first two maximas of IFFT of noise estimate, in the vicinity of zero lag , can be used to estimate the range of local SNR. However more elaborate studies need to be made for the range estimation of local SNR, i.e. the range of SNRs between which local SNR may lie.

References

[1] S. S. Stevens and J. Volkman, "Relation of Pitch to Frequency,"*American Journal of Psychology*, 1940

[2] S. Umesh, Leon Cohen, Nenad Marinovic and Douglas J. Nelson, "Scale Transform in Speech Analysis,"*IEEE Trans. on Speech and Audio Processing*, Vol. 7, No. 1, January 1999

[3] Bernard Widrow, "Adaptive Noise Cancelling: Principles and Applications,"*Proceedings of The IEEE*, Vol. 63, No. 12, December 1975

[4] Steven F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction",*IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, No. 2, April 1979

[5] Harald Gustafsson, Sven Erik Nordholm and Ingvar Claesson, "Spectral Subtraction Using Reduced Delay Convolution and Adaptive Averaging,"*IEEE Trans. on Speech and Audio Processing*, Vol. 9, No. 8, November 2001

[6] P. D. Welch, "The use of FFT for the estimation of power spectra: A method based on time averaging over short modified periodograms," *IEEE Trans. on Audio Electroacoust.*, Vol. AU-15, No. 2, pp. 70-73, June 1967

[7] L. Cohen, "The Scale Representation," *IEEE Trans. on Signal Processing*, Vol. 41, No. 12, pp. 3275-3292, December 1993

[8] R. A. Altes, "The Fourier-Mellin Transform and Mammalian Hearing," *J. Acoust. Soc. America*, Vol. 63, No. 1, pp. 174-183, January 1978

[9] S. Umesh, L. Cohen, and D. Nelson, "Frequency Warping and The Mel Scale," *IEEE Signal Processing Letters*, 2002 To Appear

[10] S. Umesh, L. Cohen, N. Marinovic and D. Nelson, "Frequency Warping in Speech," *In Proc. International Conference on Spoken Language Processing*, Philadelphia, USA, 1996

[11] Rainer Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. on Speech and Audio Processing*, Vol. 9, No. 5, July 2001

[12] Alejandro Acero, Carnegie Mellon University, "Acoustical and Environmental Robustness In Automatic Speech Recognition," *Kluwer Academic Publishers*, Phd. Dissertation Book, 1993